Università di Pisa	A.A. 2017-2018

Data Mining II **September 13, 2018**

Exercise 1 - Sequential patterns **(6 points)**

Given the input sequences listed in the table below (column 1), show for each of them **all the occurrences** of subsequences {A} → {D} and {A} → {C,D}, and finally write its total support. Repeat the exercise twice: the first time **considering no temporal constraints** (columns 2 and 4); the second time **considering min-gap = 1** (i.e. gap > 1) (columns 3 and 5). Each occurrence should be represented by its corresponding list of time stamps, e.g..: <0,2,3> = <t=0, t=2, t=3>.

column 1	*column 2*	*column 3*	*column 4*	*column 5*
	{A} → {D}		**{B} → {C,D}**	
	No constraints	*min-gap = 1*	*No constraints*	*min-gap = 1*
< {A,B,F} {C} {C,D,F} {E} {C,D} > t=0 t=1 t=2 t=3 t=4	**<0,2>, <0,4>**	**<0,2> <0,4>**	**<0,2>, <0,4>**	**<0,2>, <0,4>**
< {A,B} {C} {A,B} {C,D} > t=0 t=1 t=2 t=3	**<0,3> <2,3>**	**<0,3>**	**<0,3>, <2,3>**	**<0,3>**
< {F} {A,B,F} {A,B,C,D} {D} {E} {C} > t=0 t=1 t=2 t=3 t=4 t=5	**<1,2> <1,3> <2,3>**	**<1,3>**	**<1,2>**	**none**
< {A,F} {B,C} {A,B} {E} {D} > t=0 t=1 t=2 t=3 t=4	**<0,4> <2,4>**	**<0,4> <2,4>**	**none**	**none**
< {A,B,F} {A,C} {A,B,D} {C} {C,D} > t=0 t=1 t=2 t=3 t=4	**<0,2> <0,4> <1,2>, <1,4> <2,4>**	**<0,2> <0,4> <1,4> <2,4>**	**<0,4> <2,4>**	**<0,4> <2,4>**
Total support:	**5 (100%)**	**5 (100%)**	**4 (80%)**	**3 (60%)**

Exercise 2 - Time series / Distances (**6 points**)

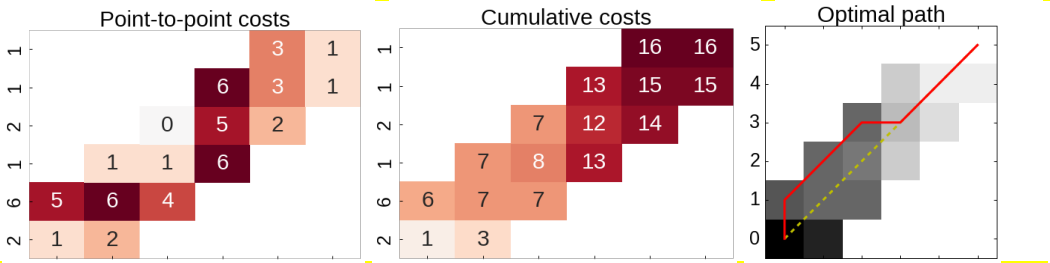
Given the following time series:

$$t = < 2, 6, 1, 2, 1, 1 >$$
$$q = < 1, 0, 2, 7, 4, 0 >$$

compute (i) their DTW, and (ii) their DTW with Sakoe-Chiba band of size r=1 (i.e. all cells at distance <= 1 from the diagonal are allowed). Show the cost matrices and the optimal paths found.

Answer:
Euclidean: sqrt(73.0) = 8.54400374532

DTW: 11



DTW r=1: 16

Exercise 3 - Analysis process & CRISP-DM **(4 points)**

Car rental companies usually pre-charge their customers of an amount of money that serves as safety buffer in case of missing fuel (in case you are expected to return the car with full tank and you don't do it) or damages to the vehicle. Car rental company X wants to make this pre-charge adaptive to the customer. In particular, the amount should be based on the age of the (main) driver and the duration of the rent. The available information is the history of past rentals, which include age of driver, duration of rent, vehicle type, how much was charged for filling the tank and how much was paid for damages. Briefly describe an analysis project aimed to help the company to realize such objective, taking as reference the CRISP-DM process.

Answer:
Realize it as a classification problem, where the inputs are are and rent duration, and the target is a discetization of total expences incurred after returning the vehicle.

Exercise 4 - Classification **(6 points)**

a) Naive Bayes (**3 points**)
Given the training set on the left, build a Naive Bayes classification model and apply it to the test set on the right.

pressure	overweight	vaccinations	class
low	yes	yes	N
medium	yes	yes	N
high	no	no	N
high	no	yes	Y
low	yes	no	Y
medium	no	no	Y
high	no	yes	Y

pressure	overweight	vaccinations	class
low	no	no	
high	yes	yes	
medium	yes	no	

Model probabilities

	Y	N
	0.57	0.43
	A | Y	A | N
high	0.50	0.33
medium	0.25	0.33
low	0.25	0.33
	B | Y	B | N
yes	0.25	0.67
no	0.75	0.33
	C | Y	C | N
yes	0.50	0.67
no	0.50	0.33

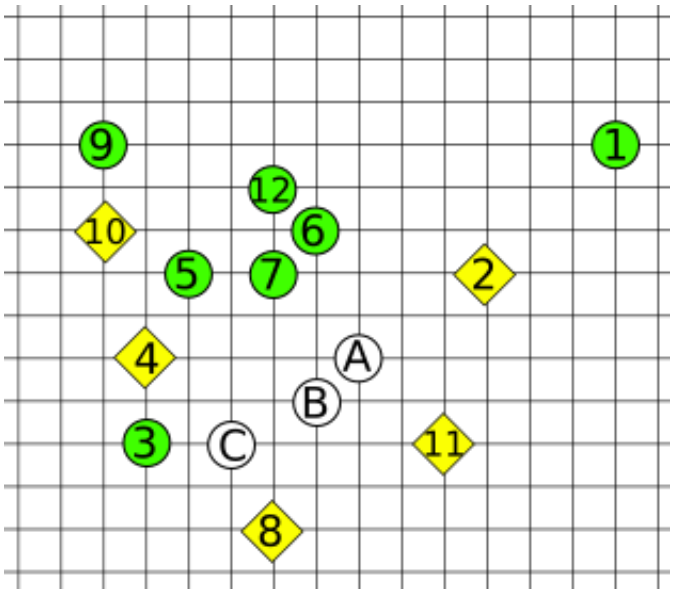
Predictions

	Class	pressure	overweight	vaccinations	
		low	no	no	
Y	0.57	0.25	0.75	0.50	0.05
N	0.43	0.33	0.33	0.33	0.02
		high	yes	yes	
Y	0.57	0.50	0.25	0.50	0.04
N	0.43	0.33	0.67	0.67	0.06
		medium	yes	no	
Y	0.57	0.25	0.25	0.50	0.02
N	0.43	0.33	0.67	0.33	0.03

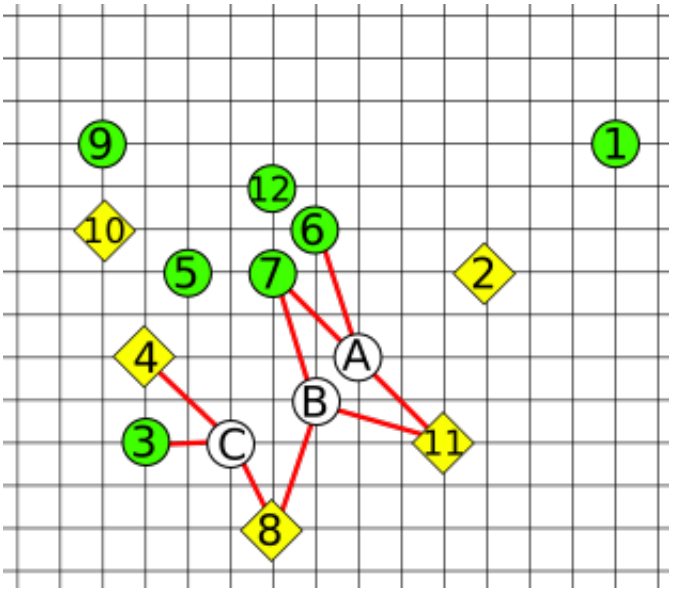
==> Output: Y, N, N

b) k-NN (**3 points**)

Given the training set on the right, composed of elements numbered from 1 to 12, and labelled as circles and diamonds, use it to classify the remaining 3 elements (letters A, B and C) using a k-NN classifier with k=3.
For each point to classify, list the points of the dataset that belong to its k-NN set.
Notice: A, B and C belong to the test set, not to the training set. Also, the Euclidean distance should be used.

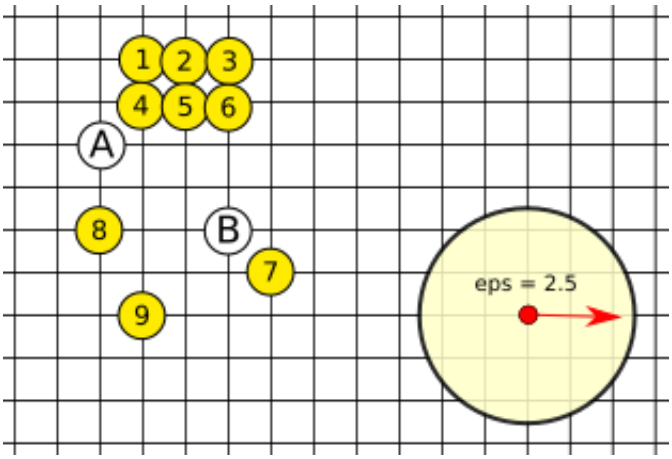


Answer:
kNN(A) = { 6, 7, 11} → CIRCLE
kNN(B) = { 7, 8, 11 } → SQUARE
kNN(C) = { 3, 4, 8 } → SQUARE



Exercise 5 - Outlier Detection **(6 points)**

Given the dataset of 11 points below (A, B, 1, 2, …, 9), consider the outlier detection problem for points A and B, adopting the following three methods:

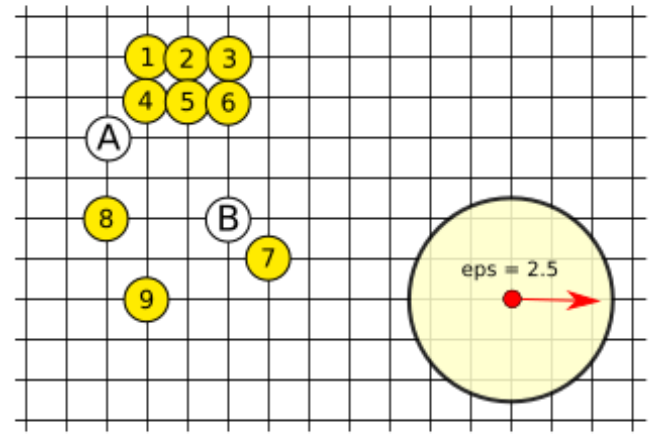


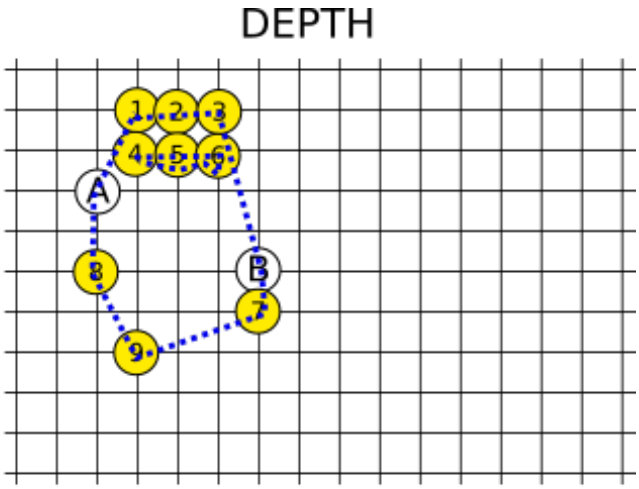
a) Distance-based: DB(ε,π) **(2 points)**
Are A and/or B outliers, if thresholds are forced to ε = 2.5 and π = 0.3? Show the density of the two points. (Notice: in computing the density of a point P, P itself should not be counted as neighbour).

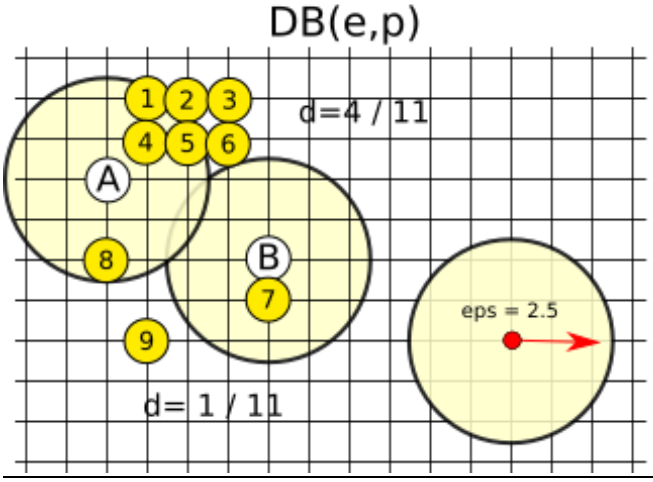
b) Density-based: LOF (3 **points**)
Compute the LOF score for points A and B by taking k=2, i.e. comparing each point with its 2-NNs (not counting the point itself). In order to simplify the calculations, the reachability-distance used by LOF can be replaced by the simple Euclidean distance.

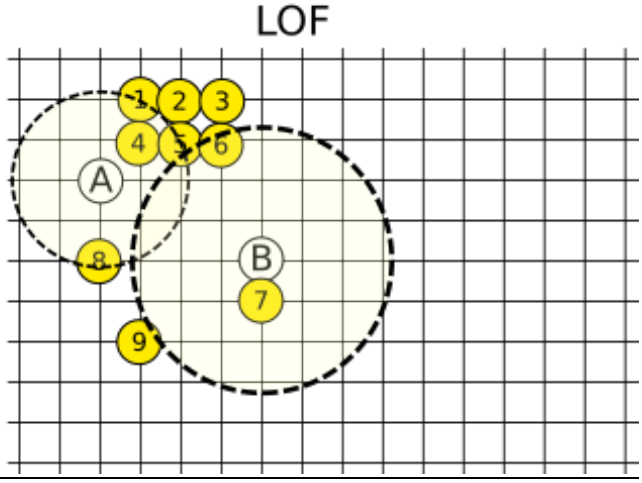
c) Depth-based (1 **points**)
Compute the depth score of all points.

Answer:









a) A → neighbours = {1,4,5,8} → d = 4/11 = 0.36 => no outlier
B → neighbours = { 7 } → d = 1/11 = 0.09 => outlier

b)

2-NN(A) = { 4, 8 } LRD(A) = 1/ [(√2 + 2)/2] = 0.586 LRD(4) = 1/ [(1 + 1)/2] = 1 LRD(8) = 1/ [(2 + √5)/2] = 0.472 **LOF(A)** = ([LRD(4) + LRD(8)]/2) / LRD(A) = [(1 + 0.472) / 2] / 0.586 = **1.256 weak outlier**	**2-NN(B) = { 6,7 }** LRD(B) = 1/ [(1 + √10) / 2] = 0.480 LRD(6) = 1/ [(1 +1) / 2] = 1 LRD(7) = 1/ [1 + √10) / 2] = 0.480 **LOF(B)** = ([LRD(6) + LRD(7)] / 2) / LRD(B) = [(1 + 0.480) / 2] / 0.480 = **1.542 weak outlier**

c) Depth 1 → 1, 2, 3, 7, 8, 9, A, B Depth 2 → 4, 5, 6

Exercise 3 - Validation **(3 points)**

ROC & AUC (**3 points**)
On a given test set below, our classification model provided the predictions and associated confidences reported on the "Predicted" column of the table. Draw the corresponding ROC curve and compute its AUC. Show the process followed to achieve that.

Record	Real Class	Predicted	
row 1	Y	N	0.83
row 2	Y	N	0.86
row 3	Y	Y	0.80
row 4	N	N	0.50
row 5	N	N	0.83
row 6	Y	N	0.88
row 7	N	N	0.96
row 8	N	N	0.71
row 9	Y	Y	0.99
row 10	Y	N	0.85

Answer:

SORTED		TPR	FPR	AUC partial
Real Class	Score	0	0	
N	0.9870911332	0	1	0
N	0.8001077686	0	2	0
Y	0.4992995541	1	2	0
N	0.2945571677	1	3	1
Y	0.1742976453	2	3	0
Y	0.1717683454	3	3	0
N	0.1526541617	3	4	3
Y	0.1360869168	4	4	0
Y	0.1156621981	5	4	0
Y	0.036508502	6	4	0
			AUC	**4**
			Normalized	**0.1666666667**